Journal of Siberian Federal University. Mathematics & Physics 2017, 10(1), 16–21

УДК 519.24

Improving the Accuracy of the Probability Density Function Estimation

Boris S. Dobronets*
Olga A. Popova†
Institute of Space and Information Technology
Siberian Federal University
Kirenskogo, 26, Krasnoyarsk, 660074
Russia

Received 03.06.2016, received in revised form 09.09.2016, accepted 10.11.2016

The paper considers the new approach to the reconstruction of the probability density function similarly the averaged shifted histogram method. An algorithm is used Richardson's extrapolation for increasing accuracy. We prove the estimates of the accuracy of the probability density function and its second derivative to choose the optimal settings for smoothing the histogram and kernel estimators and to consider the optimal choice problem of the bandwidth parameter. Presented the results of numerical experiments.

Keywords: MISE, error estimate, Richardson's extrapolation, Runge's rule, probability density functions estimation, probability density function derivatives, Numerical probabilistic analysis.
DOI: 10.17516/1997-1397-2017-10-1-16-21.

Introduction

Assessment of the probability density function and its derivatives on the empirical data is one of the most important issues in applied research [1–3]. Important to know the estimate for mathematical expectation of norm error of the constructed empirical probability density function [4].

The estimation of density derivatives has full potential for applications. This has been noted even in the first seminal papers on density estimation. This paper considers the Runge's rule application to the calculation the second derivative estimates of the probability density function. In contrast to the known methods, this approach does not require the differentiation of kernel estimates or calculations of finite differences from empirical probability density function. A detailed review of the existing methods of evaluation of derivatives and bibliography are presented in [5].

Use of estimates of the second derivatives allows to obtain realistic estimates of the mathematical expectations of l_2 error norm for the probability density function reconstruction. Knowledge of these assessments allows us to calculate the optimal bandwidth parameter h [5].

One of the first rules for practical evaluation of the error, was proposed by K. Runge in the beginning of the XX century. This rule was widely used first in the area of quadratures, then in difference methods and the finite element method. This rule is based on decomposition of the approximate solution u^h as a sum [6]

$$u^h = u + h^k v + O(h^{k+m}), \tag{1}$$

where u is the desired exact solution, v is the unknown function and h is a small discretization parameter, generally, a mesh size of the difference grid. The integer k characterizes the order

*BDobronets@yandex.ru
†OlgaArc@yandex.ru

of accuracy of the approximate solution, and $m > 0$ gives smallness of the remainder term as compared to the major error term $h^k v$. Since u and v are independent of h, for the parameter $h/2$ the following decomposition is valid:

$$u^{h/2} = u + \left(\frac{h}{2}\right)^k v + O(h^{k+m}). \tag{2}$$

Subtract it from (1) rejecting u:

$$u^h - u^{h/2} = v\left(\frac{h}{2}\right)^k (2^k - 1) + O(h^{k+m}).$$

Hence, the major error term can be determined:

$$u^{h/2} - u \approx \frac{u^h - u^{h/2}}{2^k - 1}. \tag{3}$$

Since in formula (3), the remainder term of order $O(h^{k+m})$ is rejected, it does not result in the guaranteed estimation, but with sufficiently small h it gives, in fact, an idea about the value of the numerical solution error.

Richardson's Extrapolation is a general method for generating high-accuracy results using low-order formulas. The approximation technique has an error term of predictable form [7].

Combine the two approximations in such a way that the error terms of order h^k cancel.

Multiplying (2) on 2^k and subtracting from (1) we get

$$u = \frac{2^k}{2^k - 1}u^{h/2} - \frac{1}{2^k - 1}u^h + O(h^{k+m}).$$

The article discusses a new approach for the reconstruction of the probability density function based on the empirical date. The approach is based on an approximation of the probability density function at some point with the use of windows with variable width h. Note that it has a certain similarity to the histograms in particular with a method of averaging the histograms. On the other hand, the structure and accuracy of constructed estimations are similar to kernel methods.

It is important, that using the error estimates and Richardson's extrapolation succeeded to build a refinement of solutions. Numerical example confirmed theoretical positions and showed good quality of the presented approach.

1. New approach of density estimate and mean integrated squared errors analysis

To estimate the probability density function, researchers use a variety of approaches including kernel assessment, histogram, polygrams, polygons [8]. The presented approach makes an assessment for the probability density function resembling the Histogram approach. To estimate the probability density function at some point z is used a rectangular kernel with a center at the point z with some parameter h.

Let us assume that we know the samples $\Xi = \{\xi_1, \xi_2, \ldots, \xi_N\}$ of a random variable ξ with probability density function $f(x)$ and support $[a, b]$. Let v_z^h denote the number of sample points falling in $[z - h, z + h]$. Then

$$\hat{f}^h(z) = \frac{v_z^h}{2Nh}.$$

Random variable v_z^h is recognized that v_z^h are Binomial random variables:

$$v_z \sim B(N, 2f_z^h h), \text{ where } f_z^h = \int_{z-h}^{z+h} f(x)dx/2h.$$

Expected value $\mathrm{E}[v_z^h] = Nf_z^h 2h$ or

$$\mathrm{E}[\frac{v_z}{2Nh}] = \mathrm{E}[\hat{f}^h(z)] = f_z^h$$

and variability $\mathrm{Var}[v_z] = 2Nf_z^h h(1 - 2f_z^h h)$.

Hence

$$\mathrm{Var}[\hat{f}_z^h] = \frac{\mathrm{Var}[v_z]}{(4Nh)^2} = \frac{f_z^h 2h(1-2f_z^h h)}{4h^2} = \frac{f_z^h}{2Nh} - \frac{(f_z^h)^2}{N}.$$

$$\mathrm{E}[(\hat{f}_z^h - f(z))^2] = \mathrm{E}[(\hat{f}_z^h - f_z^h)^2] + (f_z^h - f(z))^2 = Var[\hat{f}_z^h] + (f_z^h - f(z))^2.$$

$$f(x) = f(z) + f'(z)(x-z) + f''(z)(x-z)^2/2 + f^{(3)}(z)(x-z)^3/6 + f^{(4)}(\eta)(x-z)^4/24.$$

$$\mathrm{E}[\hat{f}_z^h] - f(z) = f''(z)h^2/6 + O(h^4). \tag{4}$$

Estimation at a value of $2h$ is:

$$\mathrm{E}[\hat{f}^{2h}(z)] = f_z^{2h} = \int_{z-2h}^{z+2h} f(x)dx/2h = f(z) + f''(z)4h^2/6 + O(h^4). \tag{5}$$

Estimate error is

$$\mathrm{E}[(\hat{f}_z^h - f(z))^2] = \frac{f(z)}{2Nh} - \frac{f(z)^2}{N} + (f''(z)h^2/6)^2 + O(h^6)$$

and

$$\mathrm{E}||\hat{f}^h - f||_2^2 = \frac{1}{2Nh} - \frac{||f||_2^2}{N} + \frac{||f''||^2 h^4}{36} + O(h^6). \tag{6}$$

2. Richardson's extrapolation and the second derivatives estimate

To improve the reconstruction accuracy of probability density at the point z we use the combination of kernel assessments with the parameters h and $2h$.

Let we apply Richardson's extrapolation to f_z^h and f_z^{2h}. Next, we multiply (5) on $1/4$ to subtract the result from (4) Excluding $(f''(z)h^2/6$ from (4) and (5), we get

$$f(z) = \frac{4}{3}f_z^h - \frac{1}{3}f_z^{2h} + O(h^4).$$

Let us remark that we have constructed the approximation to the function $f(z)$

$$f_{cor}^h(z) = \frac{4}{3}\hat{f}_z^h - \frac{1}{3}\hat{f}_z^{2h}, \tag{7}$$

with the accuracy $O(h^4)$.

On the other hand applying the Runge's rule we can obtain the estimate

$$f''(z) = 2(f_z^h - f_z^{2h})/h^2 + O(h^2)$$

or

$$||\hat{f}''|| = \frac{2}{h^2}||\hat{f}^h - \hat{f}^{2h}||. \tag{8}$$

3. Optimal choice problem of the bandwidth parameter h

Consider the optimal choice problem of the bandwidth parameter h from (6)

$$\mathrm{E}||\hat{f}^h - f||_2^2 = \frac{1}{2Nh} - \frac{||f||^2}{N} + \frac{||f''||^2 h^4}{36} + O(h^6) \to \min.$$

Get

$$h^* = \left(\frac{9}{2N||f''||^2}\right)^{1/5}.$$

Note that in order to find the optimum value of the parameter h^* is not necessary to know the value of the norm of the probability density function $||f||$.

In the case of kernel estimates the optimal bandwidth parameter h is determined by the formula [5]

$$h^* = \left(\frac{||K||^2}{\sigma_K^4 N||f''||^2}\right)^{1/5},$$

where

$$||K||^2 = \int K^2(x)dx$$

and

$$\sigma_K^2 = \int x^2 K(x)dx.$$

4. Numerical examples

As an example represent of improving accuracy estimates probability density function approximation errors of the sum of four uniformly distributed on [0,1] random variables.

Note that the probability density of the sum of n uniformly distributed variables is

$$p_n(x) = \frac{1}{(n-1)!}(x^{n-1} - C_n^1(x-1)^{n-1} + C_n^2(x-2)^{n-1} - \dots) \tag{9}$$

where C_n^k are binomial coefficients, and for each fixed value of the argument x sum in brackets are only for those terms for which the value of $(x-k), k = 1, 2, \dots$ nonnegative [9].

Thus, when $n = 4$ we have:

$$p(x) = \begin{cases} \frac{1}{6}x^3, & \text{if } 0 \leqslant x \leqslant 1; \\ -\frac{1}{2}x^3 + 2x^2 - 2x + \frac{2}{3}, & \text{if } 1 \leqslant x \leqslant 2; \\ \frac{1}{2}x^3 - 4x^2 + 10x - \frac{22}{3}, & \text{if } 2 \leqslant x \leqslant 3; \\ -\frac{1}{6}x^3 + 2x^2 - 8x + \frac{32}{3}, & \text{if } 3 \leqslant x \leqslant 4. \end{cases}$$

On Fig. 1 we represent numerical example with $h = 0.1$, $N = 10^4$. The solid line is exact probability density function $f(x)$, part (a): "$\circ$" is empirical probability density function $\hat{f}^h$, part (b):"$\circ$" is f_{cor}^h correction of empirical probability density function by Richardson's extrapolation, part (c): "$\circ$" is smoothing of f_{cor}^h by the nearest neighbor smoother.

Table 1. Shows a approximation of $||f''||$ by the formula (8), exact value of $||f''|| = 1.6431$.

Table 2. shows a comparison of the expected value of error before and after correction.

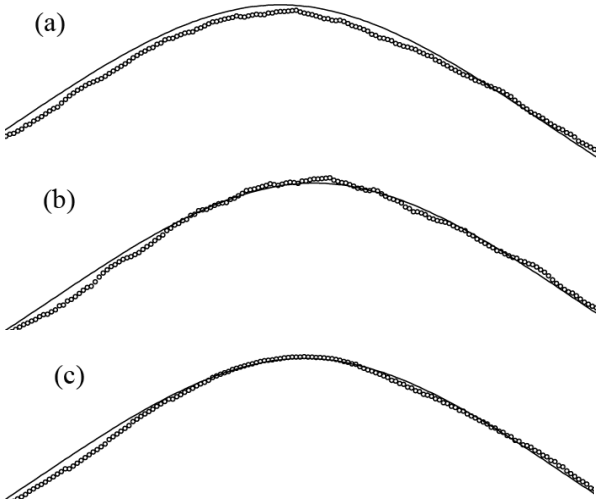

Fig. 1. (a): $\circ$ is $\hat{f}^h$, (b): $\circ$ is f^h_{cor}, (c): $\circ$ is smoothing of f^h_{cor}

Table 1. Approximation of $||f''||$

$N = 10^6$		$N = 10^4$	
h	$\|\|\hat{f}''\|\|$	h	$\|\|\hat{f}''\|\|$
0.05	1.6432	0.2	1.625162
0.1	1.6368	0.3	1.429925
0.2	1.52975	0.4	1.284578

Table 2. Correction of $\hat{f}$ by the formula (7) $N = 10^6$

h	$\|\|\hat{f}^h - f\|\|$	$\|\|f^h_{cor} - f\|\|$
0.3	0.01039296	0.00198244
0.35	0.01302775	0.00172713
0.4	0.01654605	0.00215959

Conclusion

The paper discusses the approaches to improving the accuracy the approximation of the probability density function on the empirical data. The approaches are based on the Runge's rules and Richardson's extrapolation and used to estimates of the second derivatives of the probability density function. Second derivatives estimation allowed to choose the optimal bandwidth selection for histogram and kernel estimators.

Using Richardson's extrapolation allows you to raise the calculation accuracy of the estimates for mathematical expectation of the probability density function on two orders of magnitude h.

Further development of this approach is expected in the direction of building effective smoothing procedures and bootstrap.

References

[1] B.S.Dobronets, O.A.Popova, Numerical probabilistic analysis under aleatory and epistemic uncertainty, *Reliable Computing*, **19**(2014), 274–289.

[2] B.Dobronets, O.Popova, Numerical Probabilistic Approach for Optimization Problems, Scientific Computing, Computer Arithmetic and Validated Numerics, Lecture Notes in Computer Science 9553, Springer International Publishing, Cham, 2016, 43–53.

[3] B.Dobronets, O.Popova, Chislennyi veroyatnostnyi analiz s neopredelennymi dannymi (Numerical probabilistic analysis of uncertain data), Siberian Federal University, Institute of Space and Information Technologies, Krasnoyarsk, 2014 (in Russian).

[4] O.Popova, Information approach to a posteriori error estimates of numerical modeling, *Informatization and Communication*, **2**(2016), 29–32.

[5] R.W.Scott, Multivariate density estimation: theory, practice, and visualization, John Wiley & Sons, New York, 2015.

[6] B.S.Dobronets, V.V.Shaidurov, Dvustoronnie chislennye metody (Two-sided Numerical Methods), Nauka, Novosibirsk, 1990 (in Russian).

[7] G.I.Marchuk, V.V.Shaidurov, Difference methods and their extrapolations, Springer–Verlag, New York, 1983.

[8] F.P.Tarasenko, Nonparametrics, Tomsk, TSU, 1976.

[9] S.P.Shary, Interval'nyy analiz ili metody Monte-Karlo? (Interval analysis or Monte-Carlo methods? *Vycislitel'nye tehnologii*, **12**(2007), no. 1, 103–112 (in Russian).

Повышение точности восстановления функции плотности вероятности

Борис С. Добронец
Ольга А. Попова
Институт космических и информационных технологий
Сибирский федеральный университет
Киренского, 26, Красноярск, 660074
Россия

В статье рассмотрен новый подход восстановления функции плотности вероятности, аналогичный методу осреднения сдвинутых гистограмм. Приведены алгоритмы повышения точности и вычисления второй производной, основанные на экстраполяции Ричардсона и правиле Рунге. Представлены среднеквадратичные оценки точности восстановления функции плотности вероятности и ее второй производной. Рассматривается выбор оптимальноого шага сглаживания. Представлены результаты численных экспериментов.

Ключевые слова: оценки точности, экстраполяция Ричардсона, правило Рунге, восстановление функции плотности вероятности, производные функции плотности вероятности, численный вероятностный анализ.